Predicting Optimal Cryptocurrency using Social Media Sentimental Analysis

MSc Research Project
Data Analytics

Ravi Sahal
Student ID: x21161984

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ravi Sahal
Student ID:	x21161984
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	15th December 2022
Project Title:	Predicting Optimal Cryptocurrency using Social Media Sentimental Analysis
Word Count:	6263
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	28th January 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Optimal Cryptocurrency using Social Media Sentimental Analysis

Ravi Sahal
x21161984

Abstract

Interest in cryptocurrencies has grown significantly recently on social media. Particular focus has been placed on these sorts of price adjustments. Behavioral sciences and associated academic publications have shown a significant correlation between social media and changes in cryptocurrency pricing. Smaller cryptocurrencies are particularly affected since mentions made on Twitter may have a big impact on them. Many machine learning and deep learning models were employed in recent research on cryptocurrencies to predict or anticipate the price of the coin after conducting sentiment classification. This analysis may help investors choose the best cryptocurrency. This research objective is to provide a system model for identifying the most profitable cryptocurrencies by examining data from social media platforms like Twitter. Several aspect-based sentiment models are constructed depending on the gaps that have been found. The model integrated recurrent neural network such as basic RNN and biLSTM(Bidirectional Long Short-Term Memory) with embeddings from language model (ELMo) embedding to conduct contextually-based emotional analysis on the data. As a consequence, The accuracy of the biLSTM model was 86.30% and it worked effectively when combined with the ELMo.

1 Introduction

In the current economy, cryptocurrencies are already attracting a lot of interest due to their high returns and quickest exponential development. The most valued and well-known cryptocurrency in flow right now is bitcoin, which is also the first cryptocurrency in the digital age to accrue the bulk of the market's capitalization. The exponential expansion of cryptocurrencies is driven by upward price swings of freshly listed cryptocurrencies in a very short period, inspiring many people to make an investment in them. There are two categories of investors, the first of which are those profit-seeking investors who consolidated their positions in this market and withdrew their funds over time. Another circumstance is when an actively trending price stops and begins to fall, which sends investors into chaos and as result quickly drops the number of investors. Social media platforms like Twitter are often used by individuals to convey their sentiments, ideas, and views about other persons or entities. The volatility of cryptocurrencies is gaining increasing attention on social media sites. Numerous research has shown a link between cryptocurrency returns and social media sites like Twitter (Dulău and Dulău; 2019).

1.1 Motivation and Background

The price of cryptocurrencies has been studied extensively using deep learning and machine learning based to conduct sentimental analysis on past or social media data. Natural language processing algorithms based on recurrent neural network (RNN), gated recurrent unit (GRU), long short-term memory (LSTM), and bi-directional LSTM worked well because of the precise contextual interpretation (Vo et al.; 2019). Investors could benefit from existing algorithms for determining the top 10 most advantageous cryptocurrencies, but are there any? The uniqueness of this research lies in the author's intention to develop a system that would combine user input and public opinion from multiple social media platforms, including Twitter, to recommend the top 10 most promising cryptocurrencies. When building this system, the author will utilize text filtering, text embedding, and feature engineering approaches to weed out promotional material and unearth insightful information that will improve performance.

1.2 Research Question, Objectives and Contribution

This study endeavor uses data from the top 30 trending cryptocurrencies on Twitter to suggest the top 10 cryptocurrencies to investors. Based on user tweets, deep learning algorithms were utilized to identify the mood of the tweets for each of the cryptocurrencies. A natural language processing model was selected to handle the study problem based on the examination of the literature since it was effective in recognizing and categorizing the sentiment of social media data.

RQ: To what extent contextual-based embedding ELMo can be used to develop a model with the recurrent neural network, to perform sentimental classification of social media data?

Sub RQ: How can the top 10 cryptocurrencies be generated based on sentimental classification of social media data, in this instance Twitter data, utilizing a natural language processing technique employing contextual-based embedding (ELMo Embedding) and a recurrent neural network?

The following goals are established and put into practice to answer the research question.

Research Objectives and Contribution

Obj. 1: An examination of the research on sentiment analysis of social networking website data. The outcomes of the examined literature's studies aided in understanding how sentimental analysis may be accomplished using deep learning and machine learning.
Obj. 2: Development of training data Using the Flair framework and K-means clustering. Labeling the scrapped dataset using the Flair framework and K-means for model training that was captured by scrapping the cryptocurrency-related tweets from Twitter with the help of python Twint API.
Obj. 3: Implementation, evaluation and results for sentimental classification Models.

> Sub-Obj. 3.1: Implementation, evaluation, and results for the ELMo embedding with a simple RNN Model. The result of this objective is achieved by developing a model that uses the combination of ELMo embedding and simple RNN for sentimental classification.
>
> Sub-Obj. 3.2: Implementation, evaluation, and results for the ELMo embedding with the Bidirectional LSTM Model. The result of this objective is achieved by

developing a model that uses the ELMo embedding with the Bidirectional LSTM Model for sentimental classification.

Obj. 4: Comparisons of the results of the developed model. The result of objective 4 can be done by comparing the results of the accuracy and loss values.

The report is structured as follows: Critical assessments of the literature relating to sentiment analysis utilizing various machine learning techniques are included in Section 2. The project design flow and the most recent CRISP-DM methodology approach were discussed in Section 3. Section 4 presents the implementation and evaluation of two recurrent neural network-based models with ELMo embedding. Section 5 presents the findings and discussion. Model deployment in a real-time system using a console-based application is covered in Section 6. Finally, Section 7 discusses the conclusions and future work prior acknowledgments and references.

2 Related Work

Numerous studies conducted on social media data to perform sentiment analysis using machine learning and natural language processing. The volume of posts and the expanding social media user base both significantly help with the sentiment analysis needed to forecast changes in the cryptocurrency price. The main objective of this section is to examine significant advancements in cryptocurrency forecasting and prediction systems.

2.1 Review of Sentiment analysis of Social Media Data

The research (Jianqiang and Xiaolin; 2017) was done on text pre-processing methods on Twitter data to perform sentiment analysis. The author of the study performed two feature models and four classification models on five different datasets. When different pre-processing techniques were used, the Naive Bayes (NB) and Random Forest (RF) classifiers were more accurate than Logistic Regression and support vector machine (SVM) classifiers. It has been demonstrated that removing stop words, numbers, and URLs can effectively reduce noise with little negative influence on performance. Sentiment analysis benefits from the use of negation replacement. In another paper, Numerous sentimental analyses using lexical-based methods, machine learning and deep learning methods, including SVM, Naive Bayes, LSTM and CNN, were carried out for the research project (Garg et al.; 2020). The suggested approach used a significant amount of social media data for sentiment analysis, and it was found that deep learning-based models beat machine learning and lexical-based techniques with an accuracy of 95–97%. The substantial usage of deep learning techniques can enhance the performance of this model. This suggested approach can be applied to forecast the price of cryptocurrencies in further research. The authors of the research (Hameed and Garcia-Zapirain; 2020) created a deep-learning-based model for sentiment classification of opinions which is computationally effective. The authors presented a framework that combined a single layer biLSTM model with three separate datasets from movie review websites such as MR, IMDb and SST2 to perform sentiment analysis on them. This study also compared the outcomes of other deep learning techniques and found that the proposed technique performed admirably. MR, SST2, and IMDb datasets, respectively, showed an accuracy of 80.500%, 85.780%, and 90.585% and it might be increased in the future by employing the BiGRU model.

The study (Chen et al.; 2017) discusses a novel approach that can efficiently execute sentiment analysis of various types of sentences. To solve the classification challenge, the system used a neural network and the divide and conquer strategy. The sentences in this framework are broken down into groups of sentences that can be used by the neural network to categorize once more into one-dimensional convolutional neural networks that carry out sentiment classification. The suggested method used BiLSTM-CRF to extract the target sentences, which could then be split into three categories using 1d-CNN to determine their sentimental classification. This model's accuracy ranged from 63.76 for Turkish to 73.50 for French. The performance of the model can be enhanced by using several deep-learning modules to enhance target expression identification. The authors of the paper (Rhanoui et al.; 2019) proposed a model that uses deep learning models to analyze and execute sentiment analysis on large texts. Since the lengthy text is made up of numerous sentences, it is challenging to determine the sentiment score for the entire text because it requires a lot of concentration. To analyze the lengthy text, the authors suggested a novel framework that makes use of CNN, BiLSTM, and Doc2vec embeddings. The CNN-BiLSTM model performed well with 90.66% accuracy when the author of this study compared its performance to that of CNN, LSTM, BiLSTM, and CNN-BiLSTM with word2vec and Doc2vec embeddings. Given that this was just employed with French news articles, the established approach can also be used with other languages. The research (Pimpalkar et al.; 2022) was done on social media data to do sentiment analysis and compare the findings with the other developed model. To assess and categorize the sentiments of the user reviews as positive and negative, the proposed mode of this work used the deep learning GolVe word embedding approach with multilayered BiLSTM. Using the IMDB dataset, this model was evaluated, and it did well with an accuracy rate of 93.55%. As the researcher noted, the built system might not have understood the sarcastic comments, which is why the model's precision was less than 75%. However, it can be increased with the help of new hybrid systems that use a variety of deep learning models. Most of the most recent sentiment analysis investigations use distributed word representation. This representation solely considers a word's semantic aspect and not its sentimental aspect. In the Paper (Xu et al.; 2019) To extract context-based information from the comments, the authors developed a unique model that combines TF-IDF techniques to produce weighted word vectors. These vectors were then incorporated into the BiLSTM framework. The performance aspect of this model such as accuracy score, f1 measures, recall, and precision was tested and compared with the other sentimental analysis models such as RNN, CNN, LSTM, and Naive Bayes. The proposed model achieved a precision of 91.54%, an f1 score of 92.18% and a recall value of 92.82%. The study (Aslam et al.; 2022) describes the development of a Depp learning-based model with non context-based embeddings that can analyse social media data on both emotions and sentiments. For sentiment analysis, the author combines two recurrent neural networks, the LSTM and GRU (gated recurrent unit), with a variety of deep learning algorithms, including TF-IDF, word2vec, and bag of words. and used Text2Emotion and TextBlob to analyse emotions. The suggested model performed well in terms of sentiment and emotion analysis, with an accuracy of 99% for sentiment and 92% for emotions for the tweets. This study is restricted to analysing the sentiments and feelings expressed in tweets, which can then be utilised to forecast cryptocurrency prices.

2.2 Review of Sentiment analysis Using ELMo Embeddings

The precise meaning of the given word in the text along with context cannot be captured by traditional word embedding algorithms like Word2Vec and GloVe, because they can only output a single semantic vector. To overcome this problem and to capture contextual-based sentiment scores the ELMo embedding model which is a contextual-based natural language processing model with an SVM classifier was used in a framework created for the study (Verma and Sharma; 2020). The social media data utilized in this system were scraped and passed through ELMo embeddings to provide a vector form that the SVM classifier could use to categorize the sentiment data into positive, negative, and neutral categories. Additionally, this sentiment data is utilized to investigate the correlation between Twitter data and changes in the price of bitcoin. The duration of this study was only one year, and it's highly possible that the price of bitcoin fluctuated considerably during that time. As a result, more prior data can be examined to improve the outcomes. The authors of another research (Nurifan et al.; 2019) developed a hybrid approach to do aspect-based sentiment analysis that used the ELMo embeddings with Wikipedia data as keywords extraction. This model is used to collect and process data from a restaurant review website, and the results are contrasted with the other aspects models such as Aspect Term Extraction (ATE), Aspect Keyword Extraction (AKE), Aspect Categorization (AC) and Sentiment Analysis (SA). As a result, the ELMo-Wikipedia outperformed other models, with an increase in the f1 measure of 6%. The author of the paper (Yang et al.; 2021) used the ELMo embedding model to generate the vector for the comments which were captured from a Chinese website 'JingDong' and pass this generated vector to the RNN network to get the sentimental scores. This hybrid model performed well with 88.91% of accuracy. The authors of the research(Othan and Kilimci; 2021), compared BERT, ELMo, and ULMFiT with other contextualized word embedding algorithms. These methods were used as an input layer in convolutional neural networks, recurrent neural networks, and other deep learning models (LSTMs). The investigation's findings demonstrated that, with an accuracy of 88.92, the ELMo word embedding offered the best embedding model among all others.

2.3 Review of Sentiment analysis For Cryptocurrency price prediction

For verifying the relationships between social media data and cryptocurrency prices the author of the paper (Inamdar et al.; 2019) created a system to forecast the price of the Bitcoin cryptocurrency for the following two days. For sentiment analysis, the author used RNN and LSTM and employed a random forest algorithm to estimate the price of bitcoin for the next two days. The created method accurately forecasts the price of bitcoin for the next two days, with the projected and real prices of bitcoin having the lowest MAE of 2.75 to 3.18 and the lowest RMSE of 13.70 to 15.16. The author used historical Bitcoin-related statistics to show that this technique was effective. As this was only confined to two days, the error rate can also be reduced by using numerous sources of data and longer-term predictions. A model that was developed in the paper (Dulău and Dulău; 2019) makes use of sentimental analysis for various user data sources, including Reddit and Twitter, using the Stanford CoreNLP and IBM Watson sentiment analysis systems. The top five cryptocurrencies by market share were picked for the experiment's goal in order to gather a significant amount of data. Additionally, the system used data from coinmarketcap to compare model performance to anticipated and actual values. The created algorithm analysed sentiment using more than 7000+ social data. Language processing and extensive data sets can be used to enhance the outcomes. The authors of the research, (Gurrib and Kamalov; 2021) built a system to predict the value of the Bitcoin cryptocurrency by combining the Linear Discriminant Analysis (LDA) and a Support Vector Machine model. The algorithm used data from multiple news headlines and coinmarketcap data to do sentiment analysis, and it then forecasts the direction of the Bitcoin price the following day (BTC). A trained model using data from the previous five years processed the test data, and the model did well with an accuracy of 58.5%. Since this system was solely designed to anticipate the price of Bitcoin, it cannot be used to predict the price of other cryptocurrencies. Additionally, the performance of test data can be improved by employing cutting-edge sentiment analysis algorithms like LSTM and VADER.In the Paper (Valencia et al.; 2019), the author proposed a model that extended time series forecasting through the application of machine learning and sentiment analysis approaches. The model can anticipate the performance of multi-layer predictors, SVMs, and random forest algorithms, as well as the four most popular cryptocurrencies, including Bitcoin, Ethereum, Ripple, and Litecoin. With accuracy rates of 72% for bitcoin (BTC), 44% for Ethereum (ETH), and 64% for Ripple (XRP), the project's results demonstrated that MLP performed well in the cases of the three cryptocurrencies. The SVM model performed well for Litecoin (LTC), with an accuracy rate of 66%. As the author carried out sentiment analysis using machine learning techniques. To improve the model's accuracy, deep learning techniques like LSTM and T-MLP (temporal multi-layer perceptrons) might be applied. Based on information from social media, a technique for predicting a two-hour cryptocurrency price was developed in the paper (Jain et al.; 2018). For this analysis, the author selected the cryptocurrencies Bitcoin and Litecoin due to their significant user bases and market capitalization. The author used a multivariate linear regression model to predict the price of the cryptocurrency. The developed algorithm correctly anticipated the price of Bitcoin and Litecoin with an accuracy of 44% for Bitcoin, indicating that Bitcoin's price is unaffected by sentiments in contrast to Litecoin's price, which was correctly predicted with an accuracy of 59%. By utilizing additional social factors like user popularity, user network, and other social sources, this prediction system's performance can be enhanced. The sentiment analysis of tweets to gather thoughts about cryptocurrencies and data from Google search trends to generalise public interest in cryptocurrencies are discussed in the paper (Abraham et al.; 2018). For this study, two prominent cryptocurrencies, such as Ethereum and bitcoin, were chosen based on market capitalisation. The author used the NLP Vader algorithm to do sentiment analysis on the collected tweet data and Pearson R on Google Trend data to determine a correlation between changes in cryptocurrency prices and changes in Google search trends. After using the author's model, it was possible to anticipate price change direction with more accuracy. This study demonstrates how sentiment from many tweets and Google Trend data may be used to forecast the movement of cryptocurrency prices. As the author utilised the Vader linear model to determine how cryptocurrency prices fluctuate. The employment of a complicated model rather than a linear one can boost performance. Using a novel approach that can recognise cryptocurrency-related Twitter bots, sentiment analysis was performed on the top nine major cryptocurrencies, including Bitcoin, Ethereum, XRP, Bitcoin Cash, EOS, Litecoin, Cardano, Stellar, and TRO (Kraaijeveld and De Smedt; 2020). A lexical-based system was used to perform sentiment analysis, and McDonald's financial corpus tokens and the VADER algorithm were implemented as baseline tools for the system's development. When examining the polarity scores of nine cryptocurrencies, the author found that they had a mean polarity of 0.33 for favourably skewed data at the time. This approach successfully identified the Twitter bot accounts by 1-14%. This technique can be utilised for other cryptocurrencies; however, it is only limited to 9 now. The performance of the newly established unique system for forecasting the direction of the cryptocurrency price was discussed by the authors of the paper, (Vo et al.; 2019) The system used natural language processing techniques, such as RNN, for the prediction of sentiment scores using LSTM on historical price and sentiment to carry out time series data analysis. This system examined data for a span of the previous seven days. This system performed well with a mANE value of 1.36, which measures the error rate between the predicted and real value. This technique can forecast price direction, allowing investors to base their decisions on the forecasted direction. This study found that the performance of this approach is at its lowest during periods of major upswings and downswings in ETH prices. A hybrid model for estimating the value of several cryptocurrencies, including Dash with Litecoin, Bitcoin, and Bitcoin Cash, was introduced in the paper (Parekh et al.; 2022). To assess the model performance for the combination of cryptocurrencies and provide a more precise framework for cryptocurrency price prediction, this study also examined the performance of a single cryptocurrency with several cryptocurrencies. The historical pricing for these cryptocurrencies confirms the forecasted price. The created framework improved sentimental analysis and price prediction by combining the Vader method with the LSTM model. The model accurately forecasted the price for the following 30 days using this framework, with the lowest MSE of ¡0.01, MAE of ¡0.08, and MAPE of ¡4.7 error rates. The major goal of this study was to offer a framework for predicting the price of a particular cryptocurrency. This framework can be used to analyse or predict the price of other well-known cryptocurrencies.

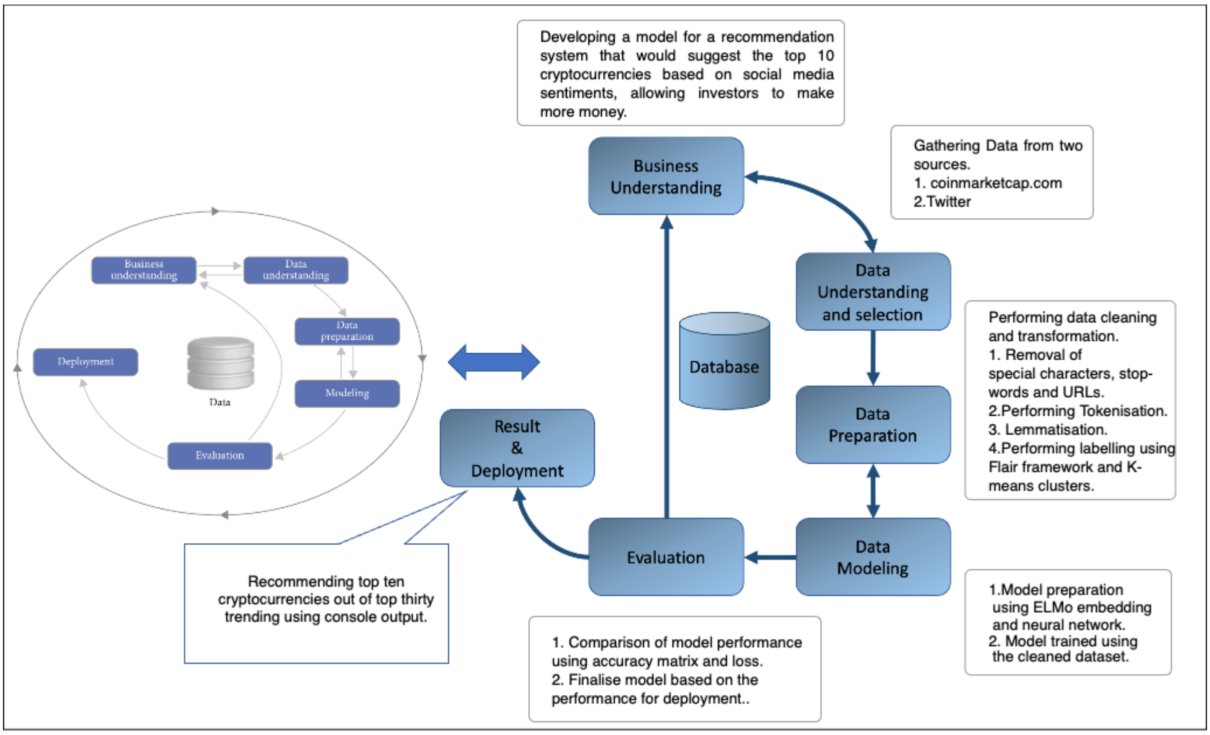
2.4 Conclusion

Therefore, considering the whole of the preceding linked to research in the disciplines, several machine learning and Natural language processing (NLP) methodologies were applied to forecast or predict the price of cryptocurrencies to investors. However, considering the previous studies, a specialized and customized recommendation system for suggesting the top 10 cryptocurrencies to investors has not yet been created. The author made the decision to create a model that can conduct contextualized based sentiment categorization to close this gap and assist investors in making more money by recommending the best cryptocurrencies.

3 Sentimental Classification-Based Methodology and Design Specifications

3.1 Sentiment Analysis for Cryptocurrency Recommendation System

In data mining-related research, KDD or CRISP-DM approaches are typically used, although in this case, CRISP-DM works best because it allows for the deployment of models at the business layer. The steps of Business Understanding and Deployment are included, which makes it different from the data mining procedures examined by KDD. The updated six stages of the CRISP-DM model for aspect-based sentiment analysis of Twitter data are described in the sections that follow (Figure 1).



Developing a model for a recommendation system that would suggest the top 10 cryptocurrencies based on social media sentiments, allowing investors to make more money.

Gathering Data from two sources.
1. coinmarketcap.com
2.Twitter

Performing data cleaning and transformation.
1. Removal of special characters, stop-words and URLs.
2.Performing Tokenisation.
3. Lemmatisation.
4.Performing labelling using Flair framework and K-means clusters.

1.Model preparation using ELMo embedding and neural network.
2. Model trained using the cleaned dataset.

1. Comparison of model performance using accuracy matrix and loss.
2. Finalise model based on the performance for deployment..

Recommending top ten cryptocurrencies out of top thirty trending using console output.

Figure 1: Modified Methodology for Sentimental Classification-Based System

3.1.1 Business understanding

The purpose of this project is to identify potential data mining-based commercial solutions for automating the work of recommending the top ten cryptocurrencies of the day to investors based on the sentimental classification of social media data. This will make it easier for a potential investor to invest in such cryptocurrencies and gain profit.

3.1.2 Data Understanding and Selection

Data has been scraped from two separate sources. The data for the top 30 trending cryptocurrencies of the week is retrieved from the trending tab of the coinmarketcap website as the initial source. Following the initial source's extraction of the top thirty cryptocurrencies' names into a JSON file, all the names are sent via a Python script to collect data from Twitter about each cryptocurrency individually and save all the tweets in CSV files with respect to their name.

3.1.3 Data Preparation

For the first data source, the data cleaning phase is completed at the time of data capture. For the second data source, data cleaning has been completed by a python script that removes unwanted columns and rows from each dataset of thirty cryptocurrencies as well as URLs, special characters, usernames, hashtags, and stop words from tweets. In this step, word lemmatization was also carried out to ensure that each word mapped to its root word. Additionally, several cleaned dataset CSVs containing clean tweets for each cryptocurrency were created.

3.1.4 Data Modelling:

The suggested model combines two deep learning models: the first is ELMo embedding, which uses clean data to train it and creates vectors for each word depending on the context of the tweet. These vectors are then fed into the second deep learning models such as Simple RNN (a fully connected RNN) and Bidirectional LSTM (biLSTM).

3.1.5 Evaluation

In this step, the performance of the proposed model is assessed using a range of factors, including the accuracy matrix and loss function.

3.1.6 Results and Deployment

The list of all the cryptocurrencies that are recommended based on the sentimental classification with the best-performed model will be displayed.

3.2 Project Design Flow

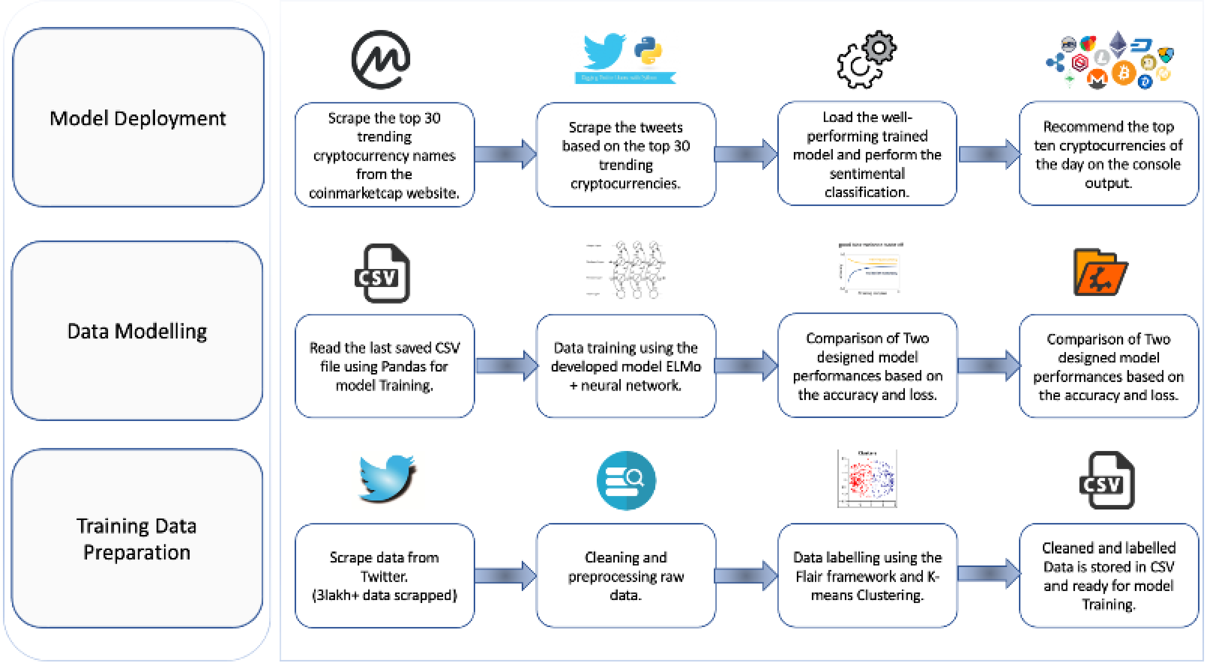

Model Deployment
Scrape the top 30 trending cryptocurrency names from the coinmarketcap website.
Scrape the tweets based on the top 30 trending cryptocurrencies.
Load the well-performing trained model and perform the sentimental classification.
Recommend the top ten cryptocurrencies of the day on the console output.
Data Modelling
Read the last saved CSV file using Pandas for model Training.
Data training using the developed model ELMo + neural network.
Comparison of Two designed model performances based on the accuracy and loss.
Comparison of Two designed model performances based on the accuracy and loss.
Training Data Preparation
Scrape data from Twitter. (3lakh+ data scrapped)
Cleaning and preprocessing raw data.
Data labelling using the Flair framework and K-means Clustering.
Cleaned and labelled Data is stored in CSV and ready for model Training.

Figure 2: Project design flow for Sentimental Classification-Based System

The recommendation system for the top ten cryptocurrencies of the day's project design process (shown in Figure 2) consists of three layers (i) the training data layer, (ii) the data modelling Layer, and (iii) the deployment layer. In the training data layer, data scraping and pre-processing are represented, In the second layer, all modelling-related data that are based on the deep learning approach are represented and the top 10 cryptocurrencies are displayed to subscribers through the console in the final layer.

4 Implementation and Evaluation of a Sentimental classification-based System

The project has been developed on a macOS operating system with a 256GB SSD and 8GB RAM. This system setup is appropriate for this study since ELMo embedding and recurrent neural networks involve a lot of processing power. Furthermore, the model and the framework were developed using Python 3.9 or later, the Jupyter Notebook, and PyCharm IDE.

4.1 Training and Testing Data Preparation

In this study, Twitter is the source of the social media data used to identify the top 10 cryptocurrencies. To collect information on cryptocurrencies from Twitter, a full-fledged framework was developed. This framework includes several scripts for data preparation, cleaning, transformation, and modelling for the datasets. For this task, a python-based selenium script was created to scrape and store the top thirty trending cryptocurrency names from the coinmarketcap.com website into a JSON file. Another Python-based script was developed that utilized Twint API and the names from the JSON file to collect public opinion data from Twitter and store them in CSVs for each of the thirty cryptocurrencies.

For the training dataset, almost 3,81,845 cryptocurrency-related tweets were successfully collected and saved into a CSV file. As the raw dataset was dirty, The Training dataset contains several undesired characters, pictures, videos, and hashtags. These must be removed at the pre-processing step since they have an influence on the calculation of the sentiment score for each tweet (Jianqiang and Xiaolin; 2017). Likewise, special characters like the question mark ("?"), exclamation point ("!"), the semicolon (";"), and "@" are deleted during pre-processing since they just add to the quantity of data needed for analysis and have no effect on polarity calculations. It was cleaned using several pre-processing methods such as stop-word removal, hashtag removal, URL removal, tokenization, and lemmatization operations with the use of nltk packages like word tokenize and WordNetLemmatizer libraries.

The pre-processed tweets must be classified considering this study involves the supervised machine learning technique. The pre-processed dataset is labeled using the state-of-the-art NLP framework Flair, which grades each document positively or negatively to indicate sentiments. The open-source natural language processing toolkit Flair is quite effective. It is mostly used for text categorization, text embedding, named entity recognition (NER), parts of speech (POS) tagging, and text classification to get insight (Ahmad and Edalati; 2022). There is no fixed guideline for the positive or negative sentiment score; this is just how the author set up the experiment. To categorises the data as either positive or negative sentiments, the author employed the k-means clustering method from scikit-learn in Python with a 2-cluster size. A total of 62,386 cleaned and transformed tweets have been categorised, with 32,445 of them being positive and 29,941 being negative.

4.2 Data Modelling

Following the labelling of each tweet with Flair, the pre-processed dataset is prepared for classification. The tweets must be converted into a numerical representation since machine learning algorithms cannot interpret plain language. In the field of natural language processing (NLP), a method known as word embeddings captures the inter-word semantics by representing individual words as vectors in a smaller-dimensional space. There are many popular Word Embedding techniques, including One Hot Encoding, TF-IDF, Glove, Word2Vec, and FastText (Rhanoui et al.; 2019). Most of the word embedding methods simply produce one vector (embedding) for each word, regardless of the context in which the text has been used, and then combine all the word's many meanings into one vector representation. Although ELMo and BERT embeddings rely on the context in which a word is used, these models provide distinct vector representations (embeddings) for a common word.

For this recommendation system, ELMo embedding is combined with a neural network to conduct sentiment categorization for Twitter data for each of the thirty cryptocurrencies. ELMo embedding, a contextual word embedding technique, is used in this research to design a model that can classify the top thirty trending cryptocurrency tweets daily. ELMo embeddings are simple to include in existing models and considerably advance the state of the art for a variety of difficult NLP issues, such as sentiment analysis, textual entailment, and question answering. To finalise the model that can identify the top ten cryptocurrencies out of the top thirty cryptocurrencies, two experiments were conducted in this research that can perform sentiment classification on the latest retrieved tweets with positive and negative sentiments. ELMo is a unique technique for representing texts in vectors or embeddings that may be fed to neural networks to carry out sentiment classification as contextual-based NLP tasks. To get a more accurate result for both experiments, the dataset is split into three sections: training data (60%) and validation data (20%) and test data (20%).

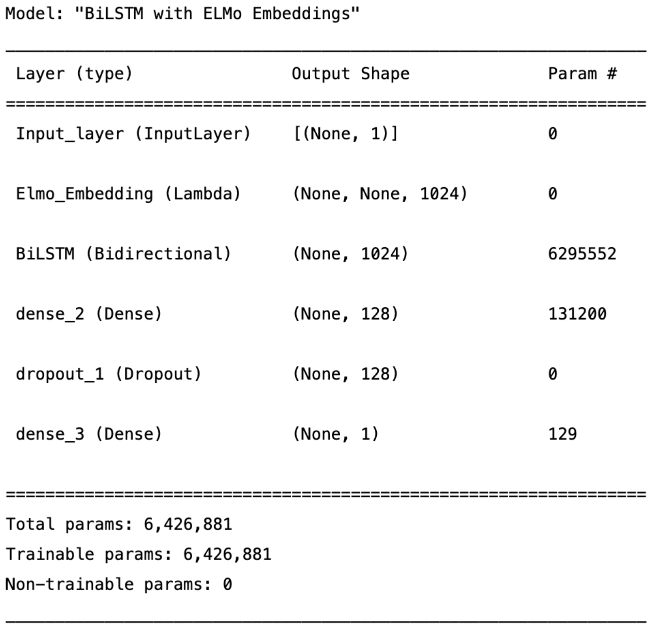
In the first experiment, the vector generated from ELMo embeddings will be used in a bidirectional LSTM model architecture's embedding layer. Bidirectional LSTM is a kind of recurrent neural network mainly used for natural language processing. In contrast to a standard LSTM, it can utilize data from both directions, and input can go both ways. It is an effective tool for identifying the correlations between words and phrases in both directions of the sequence. For the Second experiment, vector embeddings that are generated from ELMo were used along with regularisation to prevent over-fitting in a fully connected recurrent neural network. For both experiments, accuracy and loss functions were used to evaluate the model performance.

4.2.1 Implementation of ELMo embedding

To create a model that can do sentiment classification on the top 30 trending Cryptocurrency social media data. The purely character-based nature of ELMo word representations enables the network to create reliable representations for tokens that were not observed during training by using morphological cues. It creates word vectors during runtime, in contrast to other word embeddings. A system created by the author uses ELMo contextual-based embedding to produce 3-dimensional vectors for each of the data. Where the first dimension is the number of training samples, the second dimension denotes the length of the longest string in the input string list, and the third dimension denotes the length of the ELMo vector, which is 1024.

4.2.2 Implementation, evaluation, and results for the ELMo embedding with a bidirectional LSTM Model

The author built a model where the output of the embedding layer connects with the input of the biLSTM layer as per objectives 1.2 with a recurrent dropout of 20% to perform regularization on the recurrent state and a dropout of 20% for regularization on the inputs. The output of the biLSTM layer relates to a dense layer of 128 output neurons with the ReLU activation function to prevent the vanishing gradient issue and improve the computation speed. The dropout for this layer was set to be 50%. To accomplish binary classification, the output of the previous dense layer was combined with the input layer of the following dense layer, which is considered as an output layer with one neuron and a sigmoid activation function to handle binary classification. The sigmoid function is especially suitable for models whose output is a probability prediction since it can handle binary classification. It is the best choice since everything only has a chance to happen between 0 and 1. Also, as the target variable is binary, the model was built using binary cross entropy as a loss function and Adam optimizer for the quicker calculation time and it needs fewer parameter tuning and performance-evaluating parameters as an accuracy matrix.



```
Model: "BiLSTM with ELMo Embeddings"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 Input_layer (InputLayer)    [(None, 1)]               0

 Elmo_Embedding (Lambda)     (None, None, 1024)        0

 BiLSTM (Bidirectional)      (None, 1024)              6295552

 dense_2 (Dense)             (None, 128)               131200

 dropout_1 (Dropout)         (None, 128)               0

 dense_3 (Dense)             (None, 1)                 129

=================================================================
Total params: 6,426,881
Trainable params: 6,426,881
Non-trainable params: 0
_________________________________________________________________
```

Figure 3: ELMo Embedding with Bidirectional LSTM model

Evaluation: The model's loss curve and accuracy are shown in Figure 4 respectively. By identifying underfitting or overfitting issues on the training and validation data, these learning curves help to establish and improve the performance of the specified Bidirectional LSTM model. The training loss is virtually lower than the validation loss, and it steadily reduces towards the stability point on the loss curve shown in Figure 4 revealing a generalization gap. The training accuracy curve exhibits a sharp peak, followed by a moderate rise for training data and a modest increase for validation data. The model has a respectable fit overall after 7 epochs of training with a 128-batch size. The accuracy of this model was 96.28% for training data and 86.30% for testing data.

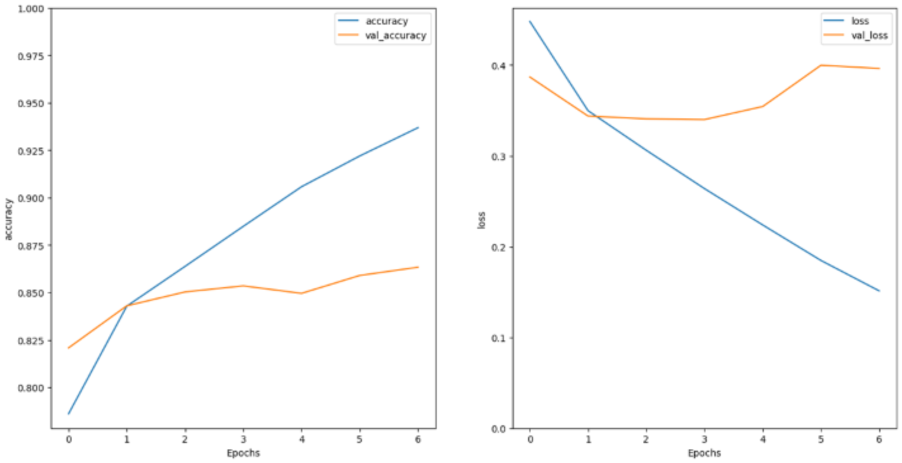
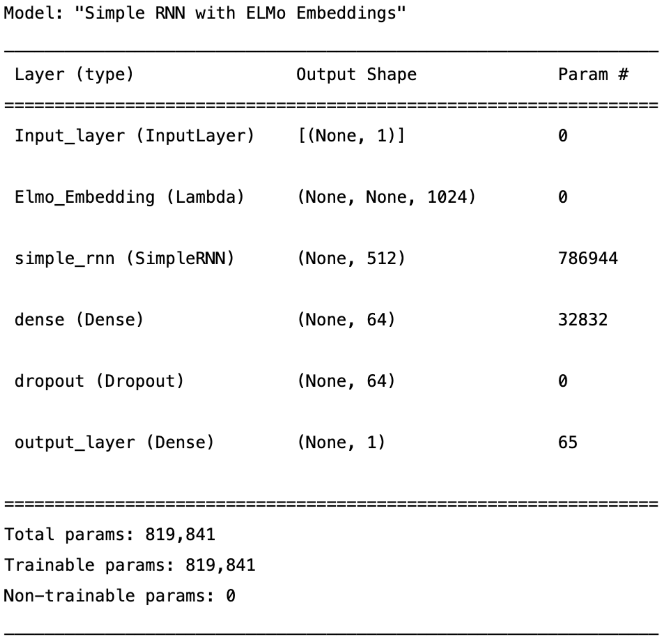


Figure 4: Accuracy and Loss learning curve for First Experiment

4.2.3 Implementation, evaluation, and results for the ELMo embedding with a simple RNN Model



```
Model: "Simple RNN with ELMo Embeddings"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 Input_layer (InputLayer)    [(None, 1)]               0

 Elmo_Embedding (Lambda)     (None, None, 1024)        0

 simple_rnn (SimpleRNN)      (None, 512)               786944

 dense (Dense)               (None, 64)                32832

 dropout (Dropout)           (None, 64)                0

 output_layer (Dense)        (None, 1)                 65

=================================================================
Total params: 819,841
Trainable params: 819,841
Non-trainable params: 0
_________________________________________________________________
```

Figure 5: ELMo Embedding with Simple RNN model

As shown in (Figure 5), For the second experiment as per objectives 1.2, the ELMo Embedding layer node to a fully connected recurrent neural layer with 512 output neurons to create a neural network. The output of the simple RNN layer is connected to another dense layer with 128 output neurons. The author then connects the preceding dense layer output to the final dense Layer, which is a representation of the neural network's output. An activation function of the first, second, and third dense layer was configured by the author to be ReLU to produce the more accurate sentiment class label (positive or negative) for each tweet. The Keras kernel regularize parameter was utilized to impose regularisation in the first dense layer to prevent overfitting. For multiclass classification tasks like this dataset, the usage of sigmoid in the output layer will be extremely suitable. The author utilized binary cross-entropy as the loss function since the dataset is labelled in a binary class. The Adam optimizer is a good alternative for optimizing the model during compilation, and the author also included accuracy as a metric for model performance.

Evaluation: A steady line in the accuracy for the model's training and validation could be seen in the learning curves (Figure 6) through the seventh epoch, which indicates that the model is not being correctly learned by the dataset as it reaches stability. However, the loss curve for train and validation data shows a little drop up to the first epoch and then stabilizes thereafter for a total of seven epochs. The model thus performed very poorly with a batch size of 128, with an accuracy of 58.80%, on the testing data.

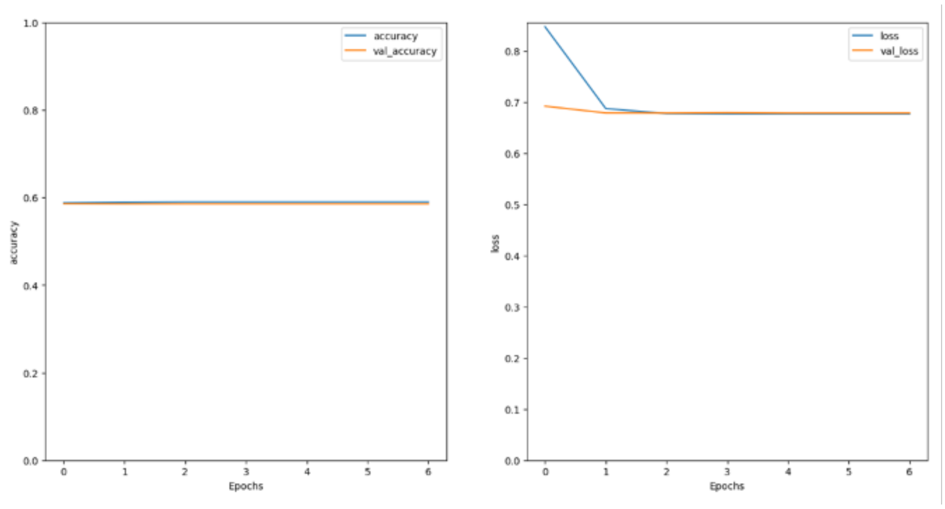


Figure 6: Accuracy and Loss learning curve for the Second Experiment

5 Results and Discussion

This section compares and discusses the performance of the bidirectional LSTM model and the simple recurrent neural language model with ELMo embedding. The primary goal of this research is to create a model that can employ context-based embedding, as ELMo used in this study with a neural network that conducts sentiment classification daily of cryptocurrency tweets 1.2. The efficiency of the developed models was compared based on the loss function and the model's accuracy. According to table 1, when compared to the Simple RNN model, the performance of the generated model that used ELMo embedding with the Bidirectional LSTM model performed well. The best model performed well, scoring 93.69% on training accuracy, 86.34% on validation accuracy, and 86.29% on test accuracy.

The second experiment, ELMo with Simple RNN, performs the worst, with a training accuracy of 59.01% and test data accuracy of 58.80%. The model using ELMo embedding with biLSTM worked effectively, as seen by the loss and accuracy learning curves for the training and validation dataset.

Table 1: Model Performance

	Accuracy and Loss		
Model	**Train Data**	**Validation Data**	**Test Data**
ELMo with bidirectional LSTM	93.69%, 0.1512	86.34%, 0.3962	86.29%
ELMo with Simple RNN	59.01%, 0.6768	58.53%, 0.6786	58.79%

6 Deployment

This section discusses the framework that has been put in place for advising investors to buy the top 10 cryptocurrencies. This framework was created with the intention of assisting investors in making more money by suggesting the day's best cryptocurrency depending on general opinion. This framework is currently being developed as a console-based application that can do the following things:

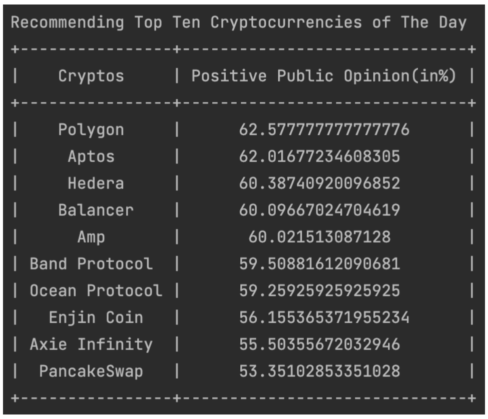
1. Extract from the coinmarketcap[1] website the top 30 trending cryptocurrency names and save them as JSON.

2. The Tweets for each of the thirty cryptocurrencies are then extracted from Twitter using the Twint API and placed in thirty distinct CSV files according to their names.

3. The thirty CSV files will undergo data cleaning and transformation operations.

4. The cleaned data will be used by the trained model, which will then conduct sentimental categorization.

5. Recommending the top ten most favourable cryptocurrencies on the console based on public opinion (Figure 7).



```
Recommending Top Ten Cryptocurrencies of The Day
+----------------+------------------------------+
|    Cryptos     | Positive Public Opinion(in%) |
+----------------+------------------------------+
|    Polygon     |      62.577777777777776      |
|     Aptos      |      62.01677234608305       |
|     Hedera     |      60.38740920096852       |
|    Balancer    |      60.09667024704619       |
|      Amp       |       60.021513087128        |
| Band Protocol  |      59.50881612090681       |
| Ocean Protocol |      59.25925925925925       |
|   Enjin Coin   |      56.155365371955234      |
| Axie Infinity  |      55.50355672032946       |
|  PancakeSwap   |      53.35102853351028       |
+----------------+------------------------------+
```

Figure 7: Top 10 Cryptocurrencies Recommended by Developed System

[1] https://coinmarketcap.com/trending-cryptocurrencies/

7 Conclusion and Future Work

Developing a recommendation system that can suggest the best cryptocurrencies to investors for them to make more money depending on the current public sentiment. Research has been done in the past to do sentiment analysis using different approaches, and most of them employ non-contextual based embeddings with the neural network because of this the performance of the system was not satisfactory since the model was not able to capture the context of the text. As a result, the research is being done in a manner that the evolving model will leverage contextual-based embedding in cooperation with a neural network to do more effective sentimental classification. The created model aims to classify the sentiment of the social media data according to the context as either positive or negative (1 or 0). To capture the text's context and produce the word vector, the developed model makes use of ELMo Embedding's strength. For classification, this word vector is later fed into the simple and bi-directional RNNs. Therefore, the first experiment worked better than the second experiment due to the usage of bi-directional LSTM, which had an accuracy of 93.69% for the training dataset and an accuracy of 86.29% for the test dataset.

The top-performing trained model was then preserved, recovered, and used to suggest the top ten cryptocurrencies out of the top thirty trending cryptocurrencies by performing the sentimental classification daily. According to recent social media comments, adopting this recommendation algorithm might help investors enhance their earnings by recommending the top 10 cryptocurrencies. This system takes a significant amount of time to do sentiment classification using the existing model, but it may be improved in the future by adding additional text-based filters and preventing the development of dense layers. In the future, the approach might be implemented on a website that also allows members to receive emails.

8 Acknowledgement

I want to sincerely thank Dr. Catherine Mulwa for her insightful and valuable advice throughout the development and execution of this research project. I am really greatful of their willingness to give their time in this manner.

References

Abraham, J., Higdon, D., Nelson, J. and Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis, *SMU Data Science Review* **1**(3): 1.

Ahmad, W. and Edalati, M. (2022). Urdu speech and text based sentiment analyzer, *arXiv preprint arXiv:2207.09163* .

Aslam, N., Rustam, F., Lee, E., Washington, P. B. and Ashraf, I. (2022). Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble lstm-gru model, *IEEE Access* **10**: 39313–39324.

Chen, T., Xu, R., He, Y. and Wang, X. (2017). Improving sentiment analysis via sentence type classification using bilstm-crf and cnn, *Expert Systems with Applications* **72**: 221–230.

Dulău, T.-M. and Dulău, M. (2019). Cryptocurrency–sentiment analysis in social media, *Acta Marisiensis. Seria Technologica* **16**(2): 1–6.

Garg, S., Panwar, D. S., Gupta, A. and Katarya, R. (2020). A literature review on sentiment analysis techniques involving social media platforms, *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, IEEE, pp. 254–259.

Gurrib, I. and Kamalov, F. (2021). Predicting bitcoin price movements using sentiment analysis: a machine learning approach, *Studies in Economics and Finance* .

Hameed, Z. and Garcia-Zapirain, B. (2020). Sentiment classification using a single-layered bilstm model, *Ieee Access* **8**: 73992–74001.

Inamdar, A., Bhagtani, A., Bhatt, S. and Shetty, P. M. (2019). Predicting cryptocurrency value using sentiment analysis, *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, IEEE, pp. 932–934.

Jain, A., Tripathi, S., Dwivedi, H. D. and Saxena, P. (2018). Forecasting price of cryptocurrencies using tweets sentiment analysis, *2018 eleventh international conference on contemporary computing (IC3)*, IEEE, pp. 1–7.

Jianqiang, Z. and Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis, *IEEE access* **5**: 2870–2879.

Kraaijeveld, O. and De Smedt, J. (2020). The predictive power of public twitter sentiment for forecasting cryptocurrency prices, *Journal of International Financial Markets, Institutions and Money* **65**: 101188.

Nurifan, F., Sarno, R. and Sungkono, K. R. (2019). Aspect based sentiment analysis for restaurant reviews using hybrid elmo-wikipedia and hybrid expanded opinion lexicon-senticircle, *International Journal of Intelligent Engineering and Systems* **12**(6): 47–58.

Othan, D. and Kilimci, Z. H. (2021). Stock market prediction with new generation deep contextualized word representations and deep learning models using user sentiments, *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, IEEE, pp. 1–6.

Parekh, R., Patel, N. P., Thakkar, N., Gupta, R., Tanwar, S., Sharma, G., Davidson, I. E. and Sharma, R. (2022). Dl-guess: Deep learning and sentiment analysis-based cryptocurrency price prediction, *IEEE Access* **10**: 35398–35409.

Pimpalkar, A. et al. (2022). Mbilstmglove: Embedding glove knowledge into the corpus using multi-layer bilstm deep learning model for social media sentiment analysis, *Expert Systems with Applications* **203**: 117581.

Rhanoui, M., Mikram, M., Yousfi, S. and Barzali, S. (2019). A cnn-bilstm model for document-level sentiment analysis, *Machine Learning and Knowledge Extraction* **1**(3): 832–847.

Valencia, F., Gómez-Espinosa, A. and Valdés-Aguirre, B. (2019). Price movement prediction of cryptocurrencies using sentiment analysis and machine learning, *Entropy* **21**(6): 589.

Verma, M. and Sharma, P. (2020). Money often costs too much: A study to investigate the effect of twitter sentiment on bitcoin price fluctuation.

Vo, A.-D., Nguyen, Q.-P. and Ock, C.-Y. (2019). Sentiment analysis of news for effective cryptocurrency price prediction, *International Journal of Knowledge Engineering* **5**(2): 47–52.

Xu, G., Meng, Y., Qiu, X., Yu, Z. and Wu, X. (2019). Sentiment analysis of comment texts based on bilstm, *Ieee Access* **7**: 51522–51532.

Yang, M., Xu, J., Luo, K. and Zhang, Y. (2021). Sentiment analysis of chinese text based on elmo-rnn model, *Journal of Physics: Conference Series*, Vol. 1748, IOP Publishing, p. 022033.